Correlation and regression vectors
The geometry of the statistics

Rastko Vukovic

May 28, 2016

Abstract

The article exposed one original interpretation of the well-known statistical correlation and regression line, in the manner of ordinary geometry and stereometry. My goal was to support the thesis in my brochure "Analysis of liberty" in Serbian, listed at the end, where I slipped that such interpretation is obvious, but it turned out not to be found anywhere in the mathematical literature or on the Internet.

1 Mean line

Given are the two series of random variables $A(a_1, a_2, \dots, a_n)$, $A'(a'_1, a'_2, \dots, a'_n)$ and series of $n \in \mathbb{N}$ units $U(1, 1, \dots, 1)$. They define three vectors $\mathbf{a} = \overrightarrow{OA}$, $\mathbf{a}' = \overrightarrow{OA}'$ and $\mathbf{u} = \overrightarrow{OU}$ in a orthogonal Cartesian system of the coordinates $O\xi_1\xi_2 \dots \xi_n$. The orthogonal projections of the points A and A' on the line OU are points M and M' respectively, as in figure 1.

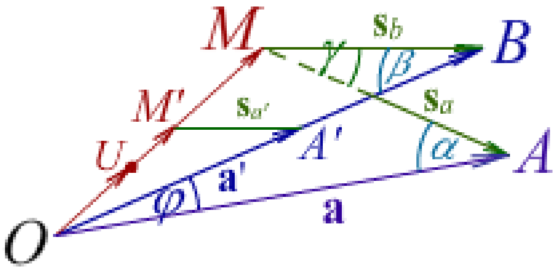

Figure 1: The right triangles OAM and $OA'M'$.

As we know, the *mean* is the average of the numbers. If we mark the means $\mu(\mathbf{a}) = \bar{a}$ and $\mu(\mathbf{a}') = \bar{a}'$, where:

$$\bar{a} = \frac{a_1 + a_2 + \dots + a_n}{n}, \quad \bar{a}' = \frac{a'_1 + a'_2 + \dots + a'_n}{n}, \tag{1}$$

then we can prove that the triangles OAM and $OA'M'$ are rectangular, with the right angles in vertexes M and M', where:

$$\mathbf{m} = \overrightarrow{OM} = \bar{a}\,\mathbf{u}, \quad \mathbf{m}' = \overrightarrow{OM}' = \bar{a}'\,\mathbf{u}. \tag{2}$$

That is the statement of the following theorem.

Theorem 1. *The triangle OAM is right-angled with $\angle M = 90°$.*

Proof. Calculate the scalar product:

$$\overrightarrow{OM} \cdot \overrightarrow{MA} = (\bar{a}, \bar{a}, \dots, \bar{a}) \cdot (a_1 - \bar{a}, a_2 - \bar{a}, \dots, a_n - \bar{a}) =$$

$$= \mu \cdot (a_1 - \bar{a}) + \mu \cdot (a_2 - \bar{a}) + \ldots + \mu \cdot (a_n - \bar{a})$$

$$= \mu \cdot (a_1 + a_2 + \ldots + a_n - n\bar{a}) = \bar{a} \cdot 0 = 0.$$

From $\overrightarrow{OM} \cdot \overrightarrow{MA} = 0$ follows $\overrightarrow{OM} \perp \overrightarrow{MA}$, that is $\angle M = 90°$. □

Let us chose a point B on a straight line OA', so that its orthogonal projection on the line OU is point M. For similar, right-angled, triangles OAM and $OA'M'$ we have the proportion $\overline{OB} : \overline{OA}' = \overline{OM} : \overline{OM}'$, then $\overrightarrow{OB} = \bar{a}\,\overrightarrow{OA}'/\bar{a}'$ or $\mathbf{b} = \bar{a}\,\mathbf{a}'/\bar{a}'$. Hence:

$$\mathbf{b} = \overrightarrow{OB} = (b_1, b_2, \dots, b_n) = \frac{\bar{a}}{\bar{a}'}\,(a_1', a_2', \dots, a_n'), \tag{3}$$

so $b_k = \frac{\bar{a}}{\bar{a}'}\,a_k'$ for all $k = 1, 2, \dots, k$. From the *mean line* OM we can see the segment AB with angles $\varphi = \angle AOB$ and $\gamma = \angle AMB$, as shown in figure 1.

Example 1. *The series $A(11, 20, 5)$ and $A'(10, 5, 3)$ present in a Cartesian system.*

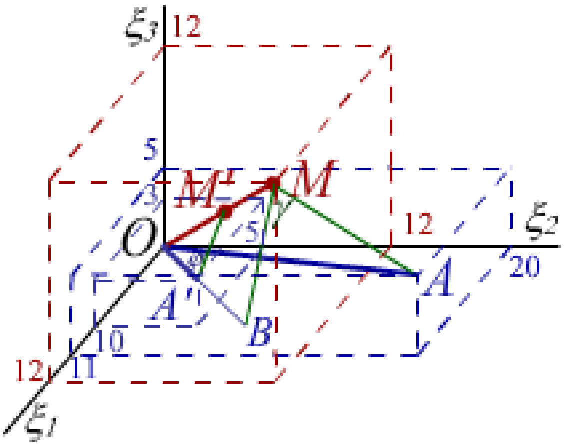

Figure 2: The series $A(11, 20, 5)$, $A'(10, 5, 3)$.

Solution. On the figure 2 we see the right triangles OAM and barely $OA'M'$. The means are $\bar{a} = 12$ and $\bar{a}' = 6$, so the point B has coordinates $B(20, 10, 6)$.

The sides of the triangle OAM are:

$$\begin{cases} \mathbf{a} = \overline{OA} = \sqrt{11^2 + 20^2 + 5^2} = \sqrt{546}, \\ \mathbf{m} = \overline{OM} = \sqrt{12^2 + 12^2 + 12^2} = \sqrt{432}, \\ \mathbf{s}_a = \overline{MA} = \sqrt{(11-12)^2 + (20-12)^2 + (5-12)^2} = \sqrt{114}, \end{cases}$$

and $\overline{OA}^2 = \overline{OM}^2 + \overline{MA}^2$, so Pythagorean theorem stands. The triangle OAM is right.

The sides of the triangle $OA'M'$ are:

$$\begin{cases} \mathbf{a}' = \overline{OA'} = \sqrt{10^2 + 5^2 + 3^2} = \sqrt{134}, \\ \mathbf{m}' = \overline{OM'} = \sqrt{6^2 + 6^2 + 6^2} = \sqrt{108}, \\ \mathbf{s}_{a'} = \overline{M'A'} = \sqrt{(10-6)^2 + (5-6)^2 + (3-6)^2} = \sqrt{26}, \end{cases}$$

and $\overline{OA}^2 = \overline{OM}^2 + \overline{MA}^2$. The triangle $OA'M'$ is right too. □

Example 2. *The series $A(-1, 2, 3)$ and $A'(4, -2, 2)$ present in a Cartesian system.*

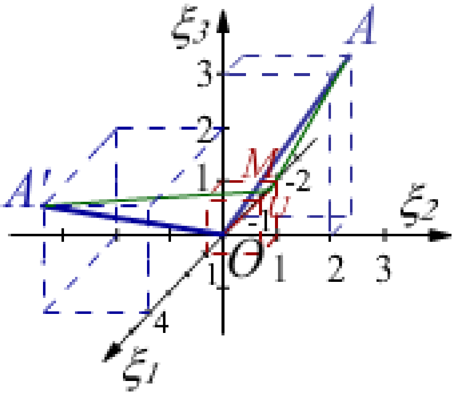

Figure 3: The series $A(-1, 2, 3)$, $A'(4, -2, 2)$.

Solution. The solution is on the figure 3. The mean of the points A and A' are the same $\bar{a} = \bar{a}' = \frac{4}{3}$, so $M(\frac{4}{3}, \frac{4}{3}, \frac{4}{3})$. For the sides of the triangle OAM we find:

$$\begin{cases} \overline{OA} = \sqrt{(-1)^2 + 2^2 + 3^2} = \sqrt{14} = \sqrt{\frac{42}{3}}, \\ \overline{OM} = \sqrt{(\frac{4}{3})^2 + (\frac{4}{3})^2 + (\frac{4}{3})^2} = \sqrt{\frac{16}{3}}, \\ \overline{MA} = \sqrt{(-1-\frac{4}{3})^2 + (2-\frac{4}{3})^2 + (3-\frac{4}{3})^2} = \sqrt{\frac{26}{3}}, \end{cases}$$

so $\overline{OA}^2 = \overline{OM}^2 + \overline{MA}^2$.

For the second triangle:

$$\begin{cases} \overline{OA'} = \sqrt{4^2 + (-2)^2 + 2^2} = \sqrt{24} = \sqrt{\frac{72}{9}}, \\ \overline{OM} = \sqrt{(\frac{4}{3})^2 + (\frac{4}{3})^2 + (\frac{4}{3})^2} = \sqrt{\frac{16}{3}}, \\ \overline{MA'} = \sqrt{(4 - \frac{4}{3})^2 + (-2 - \frac{4}{3})^2 + (2 - \frac{4}{3})^2} = \sqrt{\frac{56}{3}}, \end{cases}$$

and again $\overline{OA'}^2 = \overline{OM}^2 + \overline{MA'}^2$. □

2 Correlation triangle

The triangle ABM call *correlation triangle*. The sides of the triangle are vectors:

$$\begin{cases} \overrightarrow{MA} = \mathbf{s}_a = (a_1 - \bar{a}, a_2 - \bar{a}, \ldots, a_n - \bar{a}), \\ \overrightarrow{MB} = \mathbf{s}_b = (b_1 - \bar{b}, b_2 - \bar{b}, \ldots, b_n - \bar{b}), \\ \overrightarrow{AB} = \overrightarrow{MB} - \overrightarrow{MA} = \mathbf{s}_b - \mathbf{s}_a. \end{cases} \tag{4}$$

On the other side, from the triangle OAB we have:

$$\overrightarrow{AB} = \overrightarrow{OB} - \overrightarrow{OA} = (b_1 - a_1, b_2 - a_2, \ldots, b_n - a_n) = \mathbf{b} - \mathbf{a}, \tag{5}$$

and also $\mathbf{s}_b - \mathbf{s}_a = \mathbf{b} - \mathbf{a}$. The cosine rule for the triangles gives:

$$\begin{cases} \overline{AB}^2 = \overline{MA}^2 + \overline{MB}^2 - 2 \cdot \overline{MA} \cdot \overline{MB} \cdot \cos\gamma, \\ \overline{AB}^2 = \overline{OA}^2 + \overline{OB}^2 - 2 \cdot \overline{OA} \cdot \overline{OB} \cdot \cos\varphi. \end{cases} \tag{6}$$

From the right triangles OAM and OBM follows:

$$\begin{cases} \overline{OM} = a\sin\alpha & \overline{MA} = a\cos\alpha, \\ \overline{OB} = a\sin\alpha/\sin\beta, & \overline{MB} = a\sin\alpha\cot\beta, \end{cases} \tag{7}$$

where $a = \overline{OA}$, $\alpha = \angle OAM$, $\beta = \angle OBM$. Therefore:

$$\overline{MA}^2 + \overline{MB}^2 - 2 \cdot \overline{MA} \cdot \overline{MB} \cdot \cos\gamma = \overline{OA}^2 + \overline{OB}^2 - 2 \cdot \overline{OA} \cdot \overline{OB} \cdot \cos\varphi,$$

$$s_a^2 + s_b^2 - 2s_a s_b \cos\gamma = a^2 + b^2 - 2ab\cos\varphi,$$

$$2ab\cos\varphi - 2s_a s_b \cos\gamma = (a^2 - s_a^2) + (b^2 - s_b^2),$$

$$2(ab\cos\varphi - s_a s_b \cos\gamma) = m^2 + m^2,$$

$$ab\cos\varphi - s_a s_b \cos\gamma = m^2,$$

or by scalar product of the vectors:

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{m} \cdot \mathbf{m} + \mathbf{s}_a \cdot \mathbf{s}_b. \tag{8}$$

The marks are shown in Figure 1 and $\mathbf{m} = \overrightarrow{OM}$.

Especially, the cosine of angle between the vectors $\mathbf{s}_a$ and $\mathbf{s}_{a'}$ is:

$$\cos\gamma = \frac{\mathbf{s}_a \cdot \mathbf{s}_b}{|\mathbf{s}_a||\mathbf{s}_b|} = \frac{\sum_k (a_k - \bar{a})(b_k - \bar{b})}{\sqrt{\sum_k (a_k - \bar{a})^2}\sqrt{\sum_k \left(b_k - \bar{b}\right)^2}}. \tag{9}$$

The same result for gamma is if use $\mathbf{s}_{a'}$ instead $\mathbf{s}_b$. That is *Pearson coefficient* r of linear correlation. On the other side:

$$\cos\varphi = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|} = \frac{\sum_k a_k b_k}{\sqrt{\sum_k a_k^2}\sqrt{\sum_k b_k^2}} \tag{10}$$

It is and the cosine angle between the given vectors $\mathbf{a}$ and $\mathbf{a}'$.

Example 3. *Find the angles for the examples 1 and 2.*

Solution. In the example 1:

$$\begin{cases} \cos\gamma = \frac{-4-8+21}{\sqrt{114}\sqrt{26}} = 0.202048, & \gamma \approx 78^\circ, \\ \cos\varphi = \frac{11\cdot20+20\cdot10+5\cdot6}{\sqrt{11^2+20^2+5^2}\sqrt{20^2+10^2+6^2}} = 0.831829, & \varphi \approx 34^\circ, \end{cases} \tag{11}$$

so the angle gamma is more than twice larger than angle phi.

In the example 2:

$$\begin{cases} \cos\gamma = \frac{22/3}{\sqrt{26/3}\sqrt{56/3}} = 0.576557, & \gamma \approx 55^\circ, \\ \cos\varphi = \frac{-2}{\sqrt{14}\sqrt{24}} = -0.109109, & \varphi \approx 96^\circ. \end{cases} \tag{12}$$

Gamma is almost twice larger than phi. □

Example 4. *The first vector* $\mathbf{a}$ *has the percentage of the students that are on social welfare in 12 schools, the second* $\mathbf{a}'$ *holds the percentage of the students that use the helmet while riding the bike*[1]*.*

$\mathbf{a}$:	*50*	*11*	*2*	*19*	*26*	*73*	*81*	*51*	*11*	*2*	*19*	*25*
$\mathbf{a}'$:	*22.1*	*35.9*	*57.9*	*22.2*	*42.4*	*5.8*	*3.6*	*21.4*	*55.2*	*33.3*	*32.4*	*38.4*

Calculate angles (9) and (10).

Solution. For $\bar{a} = 370/12 = 30.833$ and $\bar{a}' = 370.6/12 = 30.883$, we get:

$\mathbf{s}_a$:	19.2	-19.8	-28.8	-11.8	-4.8	42.2	50.2	20.2	-19.8	-28.8	-11.8	-5.8
$\mathbf{s}_{a'}$:	-8.8	5.0	27.0	-8.7	11.5	-25.1	-27.3	-9.5	24.3	2.4	1.5	7.5

Then we calculate cosine (9) and (10):

$$\begin{cases} \cos\gamma = \frac{-4231.14}{\sqrt{7855.68}\sqrt{3159.68}} = -0.849266, & \gamma \approx 148^\circ, \\ \cos\varphi = \frac{7195.7}{\sqrt{19264}\sqrt{14605.0}} = 0.428992, & \varphi \approx 65^\circ. \end{cases} \tag{13}$$

The angle gamma is more than twice larger than phi, but the correlation coefficient $r = -0.85$ shows strong negative correlation. If poorer, the students use the helmet less. □

[1] From the San Jose State University, then in [1].

3 Regression line

In the correlation triangle MAB the line BC, where $C \in MA$, is perpendicular to the side MA, as you can see on the figure 4.

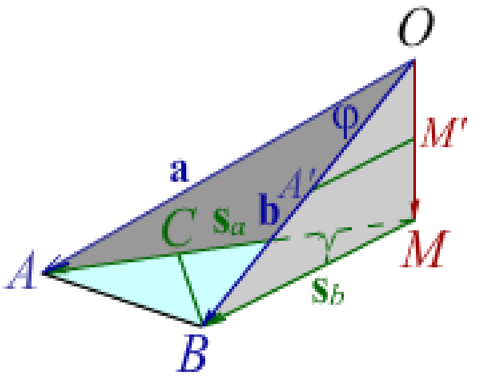

Figure 4: Line BC is perpendicular to MA.

From the previous, we know:

$$\overline{MC} = \overline{MB} \cdot \cos\gamma = s_b \cos\gamma, \quad \overline{MA} = |\mathbf{s}_a| = s_a, \tag{14}$$

which now gives:

$$x : y = \overline{MC} : \overline{MA} = \frac{s_b}{s_a} \cos\gamma = \frac{s_b}{s_a} \frac{\mathbf{s}_a \cdot \mathbf{s}_b}{|\mathbf{s}_a||\mathbf{s}_b|},$$

$$x : y = \frac{\mathbf{s}_a \cdot \mathbf{s}_b}{|\mathbf{s}_a|^2} = \frac{\sum_k (a_k - \bar{a})(b_k - \bar{b})}{\sum_k (a_k - \bar{a})^2}, \tag{15}$$

with means $\bar{b} = \bar{a}$. This is the *regression line* equation.

References

[1] Rastko Vukovic: *Analiza slobode*, inteligencija i hijerarhija (from Serbian: Liberty, Intelligence and Hierarchy), Archive.org[2], Banja Luka, May 9, 2016.

[2] Liberty: `https://archive.org/details/Sloboda`